

# Information Extraction from Mammographic Reports

**Anna Kupść,**  
**Małgorzata Marciniak,**  
**Agnieszka Mykowiecka**  
IPI PAN  
Ordonia 21  
01-237 Warszawa,  
Poland  
{aniak,mm,agn}@ipipan.waw.pl

**Jakub Piskorski**  
DFKI GmbH  
Stuhlsatzenhauseweg 3  
D-66123 Saarbuecken,  
Germany  
piskorsk@dfki.de

**Teresa Podsiadły-**  
**Marczykowska**  
IBIB  
Trojdena 4  
02-109 Warszawa,  
Poland,  
teresa@ibib.waw.pl

## Abstract

In this paper, we present an environment designed for extraction of medical data from mammographic reports. We process data collected from various Polish health care providers and transform them into attribute-value structures, according to a simplified mammographic ontology. We use a general purpose information extraction (IE) platform, SProUT, enriched with domain-specific terms. According to an adopted cascaded processing strategy the extracted pieces of information are merged externally. To the best of our knowledge, the current project is the first attempt at extracting information from Polish medical texts.

## 1 Introduction

The past few years have witnessed a growing interest in applying NLP techniques to process and understand biological and medical texts. There have been created many resources and processing tools which facilitate access to desired information. However, most of these resources are monolingual and cannot be directly reused for other languages. In this paper, we present an attempt at automatically obtaining information from Polish medical texts.

The aim of the project described in this paper is to provide a formalized description of mammography reports from various health care providers in Warsaw, Poland. As a starting point, we take a detailed hand-crafted ontology. However, to make our task realistic, we build a simplified domain model adjusted to our needs. Then, we extract partial information from the texts using SProUT (Drożdżyński et al., 2004), a general-purpose Information Extraction (IE) platform, which has been adapted to the processing of Polish (Piskorski et al., 2004). Finally, we go beyond simple IE and combine extracted

phrases together so that separate pieces of information fit our domain model.

There exist only a few processing tools for Polish, which presents an additional challenge to our enterprise. The current project is, to the best of our knowledge, the first attempt at extracting information from Polish medical texts. The similar task for English mammography reports was undertaken by (Jain and Friedman, 1997) and (Hahn et al., 2002). A Bayesian method of identifying finding's features was proposed by (Burnside et al., 2000a), while (Burnside et al., 2000b) proposed a statistical method for mapping radiology reports to BI-RADS (*Breast Imaging Reporting and Data System*) terms.

The organization of the paper is as follows: the next two sections present, respectively, the IE platform used in the project and the adopted cascaded processing strategy. Then, we describe the initial general mammographic ontology and its modification for our purposes, followed by sample extraction rules. The final processing stage, i.e., cleaning and merging of the extracted phrases, and its results are given in section 6 and 7. Section 8 contains conclusions and future work.

## 2 Grammar Development with SProUT

SProUT<sup>1</sup> is a multilingual NLP platform equipped with a set of reusable Unicode-capable online processing components for various linguistic operations, including tokenization, morphological analysis, gazetteer lookup, basic coreference resolution, etc. Since typed feature structures (TFS) are used as a uniform I/O data

---

<sup>1</sup>Shallow Text Processing with Unification and Typed Feature Structures.

```

pp :> morph & [POS Prep, SURFACE #prep,
              INFL [CASE_PREP #c]]
(morph & [POS Det, SURFACE #det,
          INFL [CASE_DET #c,
                NUMBER_DET #n,
                GENDER_DET #g]]) ?
(morph & [POS Adjective,
          INFL [CASE_ADJ #c,
                NUMBER_ADJ #n,
                GENDER_ADJ #g]])*
(morph & [POS Noun, SURFACE #n,
          INFL [CASE_NOUN #c,
                NUMBER_NOUN #n,
                GENDER_NOUN #g]])
-> phrase & [CAT pp, PREP #prep,
            CORE_NP #core_np]],
where #core_np=Append(#det, ' ', #n).

```

Figure 1: A sample SProUT rule

structure by each of these processing resources, they can be flexibly combined into a pipeline that produces several streams of linguistically annotated structures, which serve as an input for the shallow grammar interpreter, applied at the next stage. The grammar formalism in SProUT is a blend of efficient finite-state techniques and unification-based formalisms which are known to guarantee transparency and expressiveness. To be more precise, a grammar in SProUT consists of pattern/action rules, where the LHS of a rule is a regular expression over TFSs with functional operators and coreferences, representing the recognition pattern, and the RHS of a rule is a TFS specification of the output structure. Coreferences express structural identity, create dynamic value assignments, and serve as means of information transport. Functional operators are primarily utilized twofold: first, for forming the output of the rules and second, for introducing complex constraints in the rules (they can act as predicates that produce Boolean values).

A rule for recognition of prepositional phrases, presented in Fig. 1, gives an idea of the syntax of the grammar formalism. The first TFS matches a preposition. It is followed by at most one determiner, zero or more adjectives, and finally a noun. The variables **#c**, **#n**, **#g** establish coreferences expressing the agreement in case, number, and gender for the matched items (except for the preposition which solely

agrees in case with the other items). The RHS of the rule triggers the creation of a TFS of type phrase, where the surface form of the matched preposition is transported into the corresponding slot via the variable **#prep**. A value for the attribute **CORE\_NP** is created by concatenating (a call to the functional operator **Append**) the matched determiner and the noun (**#det** and **#n**). Generally, variables can be assigned arbitrarily complex TFSs as their values. All necessary types are arranged in the systems' type hierarchy, which can be modified by the user. Furthermore, grammar rules can be recursively embedded, which provides grammar writers with a context-free formalism. SProUT grammar interpreter comes with some additional functionalities, including rule prioritization and output merging mechanism (Busemann and Krieger, 2004). The former tool allows for defining a total order on a subset of grammar rules, which is used to filter out the output structures. The latter mechanism offers several techniques for merging output structures, e.g., via a sequence of unification operations. Additionally, a reference matching tool can be activated by the grammar interpreter on demand. It takes as input the output structures generated by the interpreter, potentially containing user-defined information on variant constructions (alternatives) for certain entity classes, and performs an additional pass through the text, in order to discover mentions of previously recognized entities. The variant specification is done explicitly in the grammars by defining additional attributes, e.g., **VARIANT**, on the RHS of rules, which contain a list of all variant forms.

### 3 System Architecture

One of the most successful processing strategies nowadays is a cascade of relatively simple modules aimed at solving particular subtasks. In our approach we also adopted this model and divided the extraction process into four stages: pre-processing, basic information extraction, cleaning-up the extracted data, and a final merging of data concerning one subject.

The pre-processing stage was motivated by low quality of the texts produced by physicians. There are many spelling errors (mostly lack of Polish diacritics but also other misspellings) and punctuation errors (lack of commas, periods

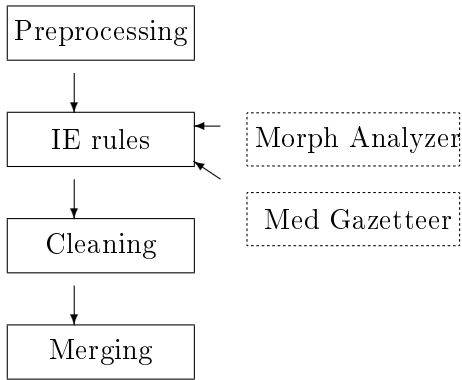


Figure 2: System architecture

or their non-standard usage). There are also many domain-specific abbreviations. Using uncorrected data would result in a severe data loss, therefore, the first step is data correction.

Polish is a language with rich inflection, so extraction from Polish medical text requires not only medical term recognition, but also recognizing various inflectional forms of the same term. Unfortunately, there exists no electronic medical lexicon for Polish. Many medical terms are also present in everyday speech and are covered by general-purpose dictionaries. Therefore, we decided to use a morphological analyser integrated with SProUT (Piskorski et al., 2004). This solution also makes it possible to use powerful SProUT rules to recognize more complicated syntactic forms, not just isolated words. Thanks to them, we can build phrases on the basis of morphological features of their elements. As only part of the medical terminology can be recognized (and inflected) by a general-purpose morphological analyser, we also employ a specialized lexicon incorporated to SProUT — the so called gazetteer.

Information extraction from unstructured texts requires a tradeoff between simplicity and extraction completeness. The more sophisticated rules we produce, the more precise information we can get; but at the same time the rules become less general. This problem begins to be extremely important when dealing with data corresponding to related features appearing freely in the text. In our application domain, this happens if we want to collect all information about a particular finding (its shape, size, contour, density, change in time, localization and so on). In addition to various permutations, pieces of relevant information can be

scattered in the document and it will be impossible to merge them locally. Therefore, we decided to process data sequentially. First, we use SProUT for extracting simple phrases — using the gazetteer and the morphological analyser we can identify all important data. Then, the results stored in an XML file are externally processed and combined into more complex AVM structures, according to the domain model presented in section 4.

The AVM structure consists of pairs of attributes and their values. A value can be atomic, another AVM structure or a list of atomic values or AVMs, see (1).

$$(1) \quad \left[ \begin{array}{l} \text{ATR1 } val1 \\ \text{ATR2 } \left[ \begin{array}{l} \text{ATR21 } val3 \\ \text{ATR22 } val4 \end{array} \right] \\ \text{ATR3 } list\ of\ values \\ \text{ATR4 } list\ of\ AVMs \end{array} \right]$$

## 4 Mammographic Ontology

The Polish mammographic ontology, a conceptual model of restricted subfield of radiology — mammography, has been based on medical literature (D’Orsi and Kopans, 1993), (Kopans et al., 1993), (Dziukowa, 1998), interviews with expert radiologist and knowledge found in the corpus (around 3000 routine free-text mammographic reports). During ontology development, concepts relevant for the domain, their properties and relations between them have been identified, domain-specific terms (with synonyms), referring to concepts and attributes have been collected. Properties of concepts have been separated into two groups: visual features of mammographic findings describing their appearance on the film, and non-visual features such as radiological diagnosis, assessment, subjective interpretation or recommendations.

The main root class in the model of mammography is the concept of Mammographic Observation. Its direct subclasses are: Breast Composition, Breast Finding and Axillary Lymph Node. They have been created from instances of subclasses of the classes Elementary Visual and Non-visual Properties of Mammographic Observation. The instances of these classes can be used to create a knowledge base of mammography, containing important mammographic le-

sions described in the literature. At the moment there are 130 classes, 342 slots and 58 instances in the model. The model has been formalized using Protégé-2000, a frame-based ontology editor developed at Stanford Medical Informatics.

In order to make our task realistic, we have simplified the above model and adjusted it to our needs. As mentioned in sec. 3, the simplified model is represented by attribute-value pairs (AVMs). The first group of attributes contains general information about an examination e.g.: an identification number, examination date, a patient identification number.

The next two attributes are descriptions of the breasts' composition. We describe separately the left and the right breasts as AVMs with the same set of attributes. In this part of the description we represent information about: types of undergone surgeries, e.g., corrective surgery, mastectomy, etc., type of tissue with detailed information about glandular tissue, its localization, density, regularity and a comparison of this information with the previous examination, and recommendations resulting from the above data.

The value of the next attribute, FINDINGS, is a list of AVMs representing the findings encountered in the mammography report. Each finding is described separately by the following attributes:

- ANAT\_CHANGE — what the finding looks like, e.g., darkness, tumor or tissue concentration,
- LOCALIZATION — an AVM with the following attributes: anatomic localization, body part, lateralization and conventional localization,
- DENSITY,
- SHAPE,
- MULTIPLICITY — necessary to represent phrases such as *liczne torbielki* 'numerous small cysts'; in this case, we create one structure describing a cyst with the value of attribute MULTIPLICITY: *numerous*,
- CONTOUR,
- CHANGE\_SIZE — an AVM with attributes describing size up to three dimensions and a measurement unit,

- WITH\_CALCIFICATION — information about accompanying calcifications (micro or macro),
- PALPABILITY — a boolean value 'yes' or 'no',
- INTERPRETATION of the finding (a cyst, cancer, an intramammary lymph node, etc.)
- DIAGNOSIS\_RTG — contains information whether the finding seems to be benign, suspicious or malignant,
- RECOMMENDATION — further examinations required,
- CHANGES\_IN\_TIME — an AVM with attributes describing the finding's changes in time.

The next group of attributes contains information about lymph nodes with their description and diagnosis.

The report often ends with a general recommendation or diagnosis which is not connected with any part of description. So we have the DIAGNOSIS and RECOMMENDATION attributes also at the main level of the AVM structure.

## 5 Sample Extraction Rules

### 5.1 Localization

A lot of information included in the mammographic examination data concerns localization of particular findings. To represent these data we defined the LOC structure containing attributes: BODY\_PART — for representing body parts (breast, armpit, etc.), L\_R — lateralization (left or right), LOC\_ANAT — anatomic localization (skin, lymph node, etc.) and LOC\_CONV — conventional name of localization, e.g., upper outer quadrant. In table 1 we show sample phrases and their representation.

The localization information is collected by several rules responsible for extracting parts of localization description. Next, the information is combined by the *loc* rule given below which creates a single complex LOC structure:

```
(2) loc:> (@seek(loc_bp) & [LOC #loc]
(token & [SURFACE -])?
@seek(loc_conv) &
[LOC_CONV #locconv, MORPH #m])?)|
```

Phrase	LOC value	
w kGZ sutka prawego ( <i>right breast – upper outer quadrant</i> )	BODY_PART L_R LOC_CONV LOC_ANAT	sutek prawy KGZ –
węzły chłonne w dołach pachowych ( <i>axillary lymph nodes in armpits</i> )	BODY_PART L_R LOC_CONV LOC_ANAT	dół pachowy prawy i lewy – węzeł chłonny

Table 1: Localization

```
(@seek(loc_conv) &
 [LOC_CONV #locconv, MORPH #m]
 @seek(loc_bp) &
 [LOC #loc, MORPH #m]) |
 @seek(loc_anat)
 & [LOC #loc, MORPH #m]
 -> [MORPH #m, LOC #loc &
 [ LOC_CONV #locconv]].
```

## 5.2 Identification of Findings

The grammar which extracts important features of findings in a mammography description consists of a set of rules, at least one for one attribute of FINDING. But there are many features which can be expressed in many ways in many orders, sometimes in several sentences.

To illustrate the problem, let us consider example (3) showing some of the many alternative ways of expressing the same information: ‘oval tumor with sharply defined margins, with dimension 10x20mm, with two calcifications — adenoma?’. In all cases, the information obtained in the extracted phase will be as in (4) but it will be differently ordered. The second processing phase is responsible for grouping these attributes into a uniform structure.

- (3)
- a. owalny, dobrze ograniczony guzek o wym. 10x20mm z dwoma zwapnieniami w jego obrębie — włókniak?
  - b. dobrze ograniczony, owalny guzek, 10x20mm. W jego obrębie widoczne są dwa zwapnienia, prawdopodobnie włókniak.
  - c. guzek o wym. 10x20mm, owalny, dobrze ograniczony, dwa zwapnienia w jego obrębie — włókniak?

- (4)
- ```
[ANAT_CHANGE zmiana]
[SHAPE owalny]
[CONTOUR dobrze_odgraniczony]
[DIAGNOSIS_RTG zmiana łagodna]
[WITH_CALCIF z zwapnieniem]
[INTERPRETATION włókniak ]
[NUM1 10, NUM2 20, DIM mm]
```

## 5.3 Finding Size

The finding size is usually given as a diameter or, in case the shape is elliptic, either two dimensions are provided or just a longer one. Sometimes, the size is specified by three dimensions. In most cases, the string corresponding to the finding size is written without spaces, e.g., ‘15x10mm’. Technically such a string is considered a single token by SProUT, therefore surface patterns are not able to separate its components, i.e., width, length and the unit. As mentioned in sec. 2, SProUT allows calls to functional operators. The system supports a handful of general operators, such as `Append` or `Boolean`, but thanks to the Java interface, the user can define other operators as well. In order to split the string, the `FindX` operator has been implemented. If the string consists of at least two sequences of digits separated by an ‘x’ and possibly followed by a sequence of letters, `FindX` produces a feature structure with attributes that store dimensions (the numbers) and an attribute for the unit (the letters).

If only one dimension is given, e.g., ‘3mm’, we check if this is a diameter (i.e., preceded by the word *średnica* ‘diameter’ or its abbreviation). In this case, we use a rule which adds the second dimension, identical with the diameter, see (2). Another functional operator, `DigitSplit`, produces a feature structure where the unit and a single number are separated.<sup>2</sup>

- (5)
- ```
split:> token & [SURFACE #s,
 TYPE number_word_first_lower]
-> #fs, where #fs=DigitSplit(#s).

finding_size:> @seek(diameter) &
 [DIAMETER yes]
 @seek(split) & [NUM1 #n1,DIM #d]
-> [NUM1 #n1, NUM2 #n1,DIM #d].
```

<sup>2</sup>Calls to other grammar rules in SProUT are executed by using the `seek` operator.

```

<?xml version='1.0' ?>
<SproutResults_XML>
<VERSION>05-2003</VERSION>
<DOCUMENT_INFO>
<DISJ>
<FS type='sprout_rule'><F name='IN'>
<FS type='*cons*'><F name='REST'>
<FS type='*null*'></FS></F><F name='FIRST'>
<FS coref='1' type='&quot;775&quot;'><
/FS></F></FS></F><F name='OUT'>
<FS type='*top*'><F name='EXAM_ID'>
<FS coref='1'></FS></F></FS></F><F name='FUNOP'>
<FS type='*cons*'><F name='REST'>
<FS type='*null*'></FS></F><F name='FIRST'>
<FS type='operator'><F name='name'>
<FS type='NumberLength'></FS></F>
<F name='args'><FS type='*cons*'>
<F name='REST'><FS type='*null*'></FS></F>
<F name='FIRST'><FS coref='1'></F></FS></F>
</FS></F></FS></F><F name='Name'>
<FS type='exam_id1'></FS></F><F name='Data'>
<FS type='Position'><F name='START'>
<FS type='0'></FS></F><F name='END'>
<FS type='0'></FS></F><F name='CSTART'>
<FS type='0'></FS></F><F name='CEND'>
<FS type='2'></FS></F></FS></F></FS>
</DISJ>
<DISJ>
...
</DOCUMENT_INFO>
</SproutResults_XML>

```

Figure 3: Sample XML output

## 6 Clean-up

We operate on the result of SProUT processing encoded in an XML file. Every recognized phrase is stored as a disjunct, with possible alternative analyses if more than one grammar rule applied to the phrase. An example of the output is given in Fig. 3.

Extracted information is stored in the `OUT` attribute in Fig. 3. Once the content of this attribute is extracted, we resolve all coreferences (e.g., the value of `EXAM_ID` is given in Fig. 3 indirectly, by reference to structure ‘1’). Then, we process the XML structures and obtain a sequence of attribute-value pairs which specify the recognized phrases, e.g., `EXAM_ID:775`.

## 7 Merging

The last processing phase identifies blocks corresponding to a finding description. We use a few

simple heuristics to group the attributes and introduce several tags to mark the beginning and the end of each block.

The two main types of blocks represent a finding and a breast’s composition description and are marked by: `up` (`uk`) — start (end) of the breast’s composition description, and `zp` (`zk`) — start (end) of a finding description.

The annotation of each report is built around the attributes representative for each block, i.e., `ANAT_CHANGE`, `INTERPRETATION` (for findings), and `BTISSUE` (for breast’s composition). Lines containing these attributes are tagged, respectively, `a_ch`, `i_ch` and `ut`. All lines with attributes which do not belong to any block (e.g., `DIAGNOSIS_RTG_LOC` or attributes starting with `BR_`) are marked as `dloc`. The last part of the report, containing general recommendations (`RECOMMENDATION`), is marked with the `rp` tag. The process of identifying blocks is repeated starting from the first line marked with `a_ch`, `i_ch` or `ut` tags. From that line we go back to the previous block’s opening or closing tag, and then go forward, trying to cover the maximal part of the report unless the `dloc` tag or attributes unique for a finding (e.g., localization, shape, size) are found. In this case, the corresponding closing tag (`uk` or `zk`) is inserted.

The pseudo-code of an algorithm for processing a sequence of attributes and grouping those corresponding to a single finding is given in Fig. 4.

Below we present sample processing results.

- (6) 775 W sutku prawym przybrodawkowo widoczny guzek o śr. 10mm z makrozwapnieniami w jego obrębie odpowiadający f-a degenerativa (zmiana łagodna). W sutku lewym w KGZ wewnątrzsutkowy węzeł chłonny.

[In the right breast in sublareolal there is a tumor of 10mm diameter with calcifications corresponding to f-a degenerativa (benign finding). In the left breast, there is an intramammary lymph node in the upper outer quadrant.]

In (7), the following information has been found and extracted:

- (7) bp  
-- EXAM\_ID:775

```

while not end of file
  find the beginning of examination
  and mark it as 'bp'
  copy one report to table TAB
  and set table TAB_TAG rows to
  'ut' if BTISSUE as found,
  to 'a_ch' for ANAT_CHANGE,
  to 'i_ch' for INTERPRETATION and to
  'dloc' for lines begining
  with BR_ and DIAGNOSIS_RTG_LOC;
skip last RECOMMENDATION lines and
  put mark 'rp' to the first of them
skip identification information;
checkpoint=beginning of examination;
while not end-of-examination and not 'rp'
  find next 'ut', 'a_ch' or 'i_ch';
  go back to the nearest block boundary
  ('zk', 'uk', 'dloc', 'bp');
  check wheather unique attributes are
  not repeated and correct boundary;
  mark the boundary as 'up' or 'zp'
  go forward while tag notequal to
  'ut', 'a_ch', 'dloc', 'rp' or
  'i_ch' (only if started from 'i_ch');
  check wheather unique attributes are
  not repeated and correct boundary;
  mark the boundary as 'uk' or 'zk';
  checkpoint= last boundary+1;
print out the results

```

Figure 4: Identifying findings

```

zp
-- LOC|BODY_PART:sutek
||LOC|LOC_CONV:ok. brodawki sutkowej
||LOC|L_R:prawy
-- ANAT_CHANGE:guzek||MULT:singular
-- DIM:mm||NUM1:10||NUM2:10
-- C_MULT:plural
||WITH_CALCIF:makrozwapnienie
-- INTERPRETATION:f-a degenerativa
-- DIAGNOSIS_RTG:zmiana_lagodna
zk
zp
-- LOC|BODY_PART:sutek
||LOC|LOC_CONV:loc_KGZ||LOC|L_R:lewy
-- INTERPRETATION:
wewnatrzsutkowy węzeł chłonny
zk
bk

```

In (7), identifying a new localization (the attribute unique for a finding) is a good criterion for separating findings' descriptions. However, in some reports this strategy leads to wrong segmentations. In (8), for the second finding, only

the interpretation is given. As its localization occurs after the finding and there is no interpretation for the first finding, 'intramammary lymph node' is classified as an interpretation of 'density'.

(8) 123 Sutek prawy – w kwadrancie górnym zagęszczenie dobrze wysyczone o średnicy około 20 mm i zatartych granicach. Wymaga ona dalszej diagnostyki – konieczne wykonanie badania USG i PCI. Wewnatrzsutkowy węzeł chłonny w kwadrancie górno-zewnętrzny sutka lewego.

[The right breast – in the upper outer quadrant there is a high density finding of about 20 mm diameter and obscured margins. Requires further examination – USG and biopsy compulsory. An intramammary lymph node in the upper outer quadrant.]

```

(9) bp
-- EXAM_ID:123
zp
-- LOC|BODY_PART:sutek
||LOC|LOC_CONV:loc_KGZ||LOC|L_R:prawy
-- ANAT_CHANGE:zagęszczenie
||MULT:singular
-- SATURATION:dobrze wysyczone
-- DIM:mm||NUM1:20||NUM2:20
-- CONTOUR:zatrzeć zarysy
-- RECOMMENDATION:USG_PCI||TIME:unknown
-- INTERPRETATION:
Wewnatrzsutkowy węzeł chłonny
zk
-- LOC|BODY_PART:sutek
||LOC|LOC_CONV:loc_KGZ||LOC|L_R:lewy
bk

```

Currently, we have started evaluation of the annotation algorithm. In our preliminary evaluation of 70 reports, we obtained 80.8% precision and 86.7% recall in annotating the beginning of a finding description. At the moment, we have identified the following main reasons of detected errors: 1) coordination — some elements of conjoined phrases are not repeated and in most cases this results in identifying only one of the conjoined elements; 2) negated phrases — not all forms of negation have been captured by shallow extraction rules, which caused opposite interpretations; 3) paraphrases — different

ways of expressing the same concept disallowed its full recognition.

The full evaluation will be presented in the final version of the paper.

## 8 Conclusions and Future Work

The paper presents a combined approach to information extraction from mammographic examination reports. IE from brief and compacted texts, meant originally as notes for other physicians, turned out to be a quite challenging task. Main processing issues were caused by the lack of a clear document structure, style differences between various physicians, multiple paraphrases, tendency to make texts very short and intensive use of individual abbreviations. Another problem was a great discrepancy between the general mammographic ontology, developed mainly on the basis of medical knowledge, and formulations found in the reports: very often they could not be directly translated into ontology concepts as statements used in reports were unclear, incomplete and ambiguous.

In the paper, we divided the extraction process into three steps. As the information we needed to extract was often scattered in the reports, we decided to first extract smaller pieces of information and, then, combine them externally into one attribute-value structure. This solution turned out to be quite successful in collecting the data but still a lot of problems have to be resolved. The grouping procedures implemented so far are based on rather simple heuristics which cannot capture more complicated cases. For example, elliptic references to previous findings, as in ‘There are several changes in the left breast, *the greatest* of 2cm size’ or relative phrases such as ‘*a similar finding*’ will remain uninterpreted. Hence, the next step will be to enhance grouping rules so that more complex cases are covered and provide full evaluation.

We also plan to incorporate an inference mechanism. This would allow for filling in data missing from the reports but which can be inferred based on general medical knowledge. After the amendments, data will be entered to a database where they can be further analysed.

## References

- E. Burnside, D. Rubin, and R. Shachter. 2000a. A Bayesian network for mammography. In *Proceedings of the American Medical Informatics Association Symposium*, pages 16–110.
- E. Burnside, H. Strasberg, and D. Rabin. 2000b. Automated indexing of mammography using linear least squares fit. In *CARS 2000 International Conference on Computer Assisted Radiology and Surgery, San Francisco, CA*.
- S. Busemann and H.-U. Krieger. 2004. Resources and techniques for multilingual information extraction. In *Proceedings of LREC 2004, Lisbon, Portugal*.
- C. J. D’Orsi and D.B. Kopans. 1993. Mammographic feature analysis. *Seminars in Roentgenology*, 28:204–230.
- W. Drożdżyński, H-U. Krieger, J. Piskorski, U. Schäfer, and F. Xu. 2004. Shallow processing with unification and typed feature structures - foundations and applications. *German AI Journal KI-Zeitschrift*, 01/04. Gesellschaft für Informatik e.V.
- J. Dziukowa. 1998. *Mammografia w Diagnostyce Raka Sutka*. Scientific Publ. Co., Warszawa.
- U. Hahn, M. Romacker, and S. Schultz. 2002. MEDSYNDIKATE— a natural language system for the extraction of medical information from findings reports. *International Journal of Medical Informatics*, pages 63–74.
- N. L. Jain and Carol Friedman. 1997. Identification of findings suspicious for breast cancer based on natural language processing of mammogram reports. In *Proceedings of the American Medical Informatics Association Annual Fall Symposium*, pages 829–833.
- D. B. Kopans, C.J. D’Orsi, D.D Adler, and al. 1993. Breast imaging reporting and data system (BI-RADS). In *American College of Radiology*.
- J. Piskorski, P. Homola, M. Marciniak, A. Mykowiecka, A. Przepiórkowski, and M. Woliński. 2004. Information Extraction for Polish Using the SProUT Platform. In *Intelligent Information Processing and Web Mining. Proceedings of the IIS’04 Conference, Zakopane*. Springer.